ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

Speech Emotion Recognition Using Machine Learning

Tanvi Shirbhate[1], Devashish Deshmukh[2], Chetan Rajurkar[3], Sayali Sagane[4], Prof. (Dr) A. W. Burange[5]

[1,2,3,4] *Undergraduate Student, Prof. Ram Meghe Institute of Technology and Research Badnera, (MS), India*

[5] *Assistant Professor, Prof. Ram Meghe Institute of Technology and Research Badnera, (MS), India*

Abstract: *Language is the most important medium of communication. Emotions play an important role in human life. Recognizing emotion in speech is both important and challenging because we are dealing with human-computer interaction. Speech Emotion Recognition (SER) has many applications, and a lot of research has focused on this interest in recent years. Speech Emotion Recognition (SER) has become an important collaboration at the intersection of music processing and machine learning. The goal of the system is to identify and classify emotions in speech, leading to human-computer applications, psychological assessments, and other tasks. The framework of the report includes preliminary data, feature extraction, model selection, training and evaluation. Our work aims to use machine learning techniques to improve speech recognition skills. It uses different models to recognize emotions and identifies emotions such as happiness, sadness, anger, surprise, etc. The machine converts human voice signals into waveforms executes its programs and finally expresses emotions. The data are speech samples and features are extracted from the speech samples using the librosa package. The RAVDESS dataset used as the test dataset shows that the accuracy of all classifiers reaches 86% for our dataset.*

Keywords: Speech Emotion Recognition (SER), Machine Learning, classifier, Feature extraction, RAVDESS Dataset, Mel-frequency cepstral coefficients (MFCCs), Librosa Package, etc.

I. INTRODUCTION

Speech Emotion Recognition (SER) is defined as the process of extracting a speaker's emotions from his or her speech. It is widely used in the service centre to understand the reaction of people involved in the call to the customer, in the car to understand the driver's mind to prevent accidents, as a diagnostic tool to detect various disorders of patient's medical services, in E-tutoring story-telling applications to adapt according to the mood of the listeners, etc. [5]. In this study, for feature extraction different algorithms such as MFCC (Mel Frequency Cepstral Coefficient) and DWT (Discrete Wavelet Transform) are used, with MLP classifier and various Python libraries are used to successfully classify the emotions [8]. To communicate effectively with humans, machines need to understand emotions in speech. For this reason, it is necessary to create machines that can think about communicating well and, like humans, can recognize the language of communication. The goal of creating machines to interpret paralinguistic information (such as emotions) facilitates human-computer interaction and helps interaction become clearer and better [10].

II. MACHINE LEARNING

Machine learning is a widely used tool to predict or classify data to support decision-making. Machine learning algorithms are trained on historical data to learn patterns and make predictions. However,

 DOI: 10.5281/zenodo.11049046

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

just creating a model is not enough. Optimization and tuning are important to ensure accuracy. This includes tuning hyperparameters to achieve optimal performance. The model analyzes the decision-making process as it is repeated over examples. It uses these learning models to make predictions when it receives new data. By optimizing models using the latest normalization methods, machine learning can adapt to new models and improve results over time.

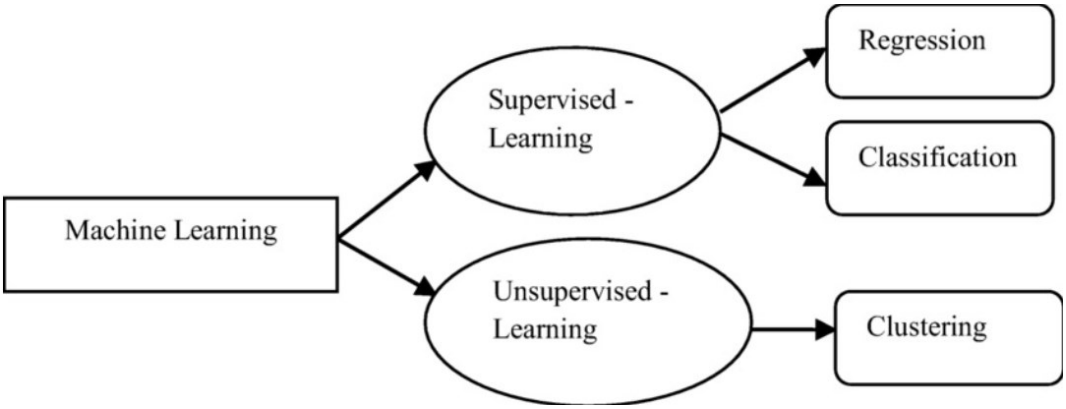


Figure 1: Types of Machine Learning

1. **Supervised Learning:** Utilizes labelled data to train machines, teaching them to predict outputs accurately. It encompasses regression, for estimating continuous values, and classification, for grouping outputs into classes.
   Regression: Models target prediction values based on independent variables, commonly used for forecasting relationships between variables.
   Classification: Groups output into classes based on discrete or categorical data, facilitating tasks like categorizing house prices in real estate markets.

2. **Unsupervised Learning:** Discovers patterns from unlabeled data, identifying groupings (clustering) and rules (association) autonomously.

III. LITERATURE REVIEW

Utkarsh Garg et.al (Nov 2020). In the Speech Emotion Recognition (SER) within the Human-Computer Interaction domain, significant emphasis is placed on the extraction and combination of audio features to create feature vectors, crucial for the system's ability to discern emotions from speech. The study leverages Mel-Frequency Cepstral Coefficients (MFCC), Mel Spectrogram features, and Chroma features, each bringing unique characteristics that capture the essence of speech with human emotions. MFCCs are instrumental for their capacity to encapsulate voice timbre, Mel Spectrogram features for translating frequencies into a scale that mirrors human auditory perception, and Chroma features for highlighting the harmonic and melodic aspects across different pitch classes. These features are ingeniously combined in various ways to form feature vectors, serving as the foundation for the subsequent classification tasks. This approach underlines the critical role of feature selection

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

and combination in enhancing the efficacy of SER systems, providing a pathway for the identification of the most potent combinations for emotion recognition [1].

Anurish Gangrade et.al (July 2022). In advancing the domain of speech emotion recognition (SER), this study introduces a groundbreaking approach by integrating Deep Belief Networks (DBNs) with an enhanced Convolutional Neural Network (CNN) model for the refined extraction of emotional cues from speech signals. The method employs a 9-layer DBN to process extensive sequences of speech frames, generating a detailed high-dimensional feature set. This is complemented by a sophisticated 9-layer CNN architecture, specifically tailored with one-dimensional convolutional layers to adeptly handle the nuances of speech data.
Evaluated on a diverse set of five emotions across male and female speakers, the system demonstrated a remarkable emotion recognition rate of 89.00%, surpassing traditional methods by approximately 14%. This significant improvement underscores the efficacy of combining deep learning architectures for feature extraction in enhancing the accuracy of SER systems, marking a notable advancement in the field [2].

Girija Deshmukh et.al (March 2019). In exploring the domain of emotion recognition within speech signals, a pivotal study leverages the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, dividing it into subsets for training and testing purposes. This research underscores the significance of Mel-frequency cepstral coefficients (MFCCs) and energy levels as fundamental features for capturing the nuanced vocal tract dynamics and intensity variations associated with different emotional states in speech.
By meticulously extracting these feature vectors from both acted and naturally occurring speech across varied emotional spectrums namely neutrality, anger, fear, and sadness the study sets a precedent for the development of a classifier model. This model, trained on the intricate patterns distilled from the MFCC and energy features, showcases a sophisticated approach to accurately discerning and categorizing emotions in speech. This methodology not only enriches the literature on audio signal processing but also elevates the precision in emotional recognition, offering profound implications for enhancing human-computer interaction through more empathetic and responsive AI systems [3].

S. G. Shaila et.al (May 2023). A comprehensive investigation into Speech Emotion Recognition (SER) within the realm of Human-Computer Interaction (HCI), focusing on the utilization of the RAVDESS dataset and machine learning methodologies for emotion classification. Notably, the study underscores the growing importance of understanding emotions in HCI, recognizing them as crucial cues for effective user interaction. By leveraging SER techniques, the research aims to discern emotional states from audio signals, thereby enhancing the capabilities of HCI systems to interpret and respond to human emotions. The integration of SER signals with Brain-Computer Interfaces (BCI) further broadens the scope of the study, highlighting its interdisciplinary nature and potential applications beyond traditional HCI contexts.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

Through a rigorous methodological approach encompassing noise reduction, feature extraction, and machine learning-based classification using models such as Random Forest, Multilayer Perceptron, Support Vector Machine, Convolutional Neural Network, and Decision Tree, the study seeks to achieve robust emotion classification accuracy. These findings contribute significantly to the literature on emotion recognition in HCI, providing insights into effective methodologies and machine-learning techniques for enhancing human-computer interaction experiences [4].

T. Kishore Kumar (July 2019). A novel approach to recognizing stressed emotions in speech signals through the integration of the Teager Energy Operator (TEO) and Mel Frequency Cepstral Coefficients (MFCC), termed Teager-MFCC (T-MFCC). TEO, a nonlinear signal processing technique designed to capture the dynamic properties of speech, particularly those associated with stress, is combined with MFCC, a widely used method for extracting spectral features from speech signals. This fusion of features aims to provide a comprehensive representation of speech signals, emphasizing both nonlinear dynamics and spectral characteristics relevant to stress.
Following feature extraction using T-MFCC, classification is performed using a Gaussian Mixture Model (GMM) to distinguish between different stressed emotions and neutral speech. The study evaluates the performance of the proposed T-MFCC method compared to traditional MFCC-based approaches, demonstrating its superior effectiveness in recognizing stressed emotions from speech signals. This research contributes to the literature by introducing a novel feature fusion technique that enhances the accuracy of emotion recognition in speech-processing tasks [5].

Chen Caihua (July 2019). Encapsulates a comprehensive investigation into speech-based emotion recognition, focusing on the integration of multi-modal fusion techniques to enhance recognition accuracy. The study delves into key aspects such as speech signal pre-processing, feature extraction, fusion strategy, fusion method, and emotion recognition classification. By constructing a theoretical model for emotional recognition and conducting research on feature fusion and classification algorithms, the paper aims to advance the understanding and development of robust emotion recognition systems.
Utilizing Support Vector Machines (SVM) for speech signal processing, the research meticulously details methods for emotional feature extraction, emphasizing the nuanced analysis of speech signals to capture the intricacies of emotional expression. The validation of extracted features through recognition results underscores the effectiveness of the proposed approach in improving emotion recognition accuracy. Overall, the study contributes valuable insights to the field of emotion recognition by employing a comprehensive multi-modal fusion approach and elucidating the intricate processes involved in accurately recognizing emotions from speech signals [6].

Husbaan I. Attar et.al (May 2022). The excerpt presents a chapter dedicated to advancing emotion recognition technology within Affective Computing by proposing a novel approach to emotion recognition from continuous speech. Unlike the prevalent focus on recognizing emotions from isolated short sentences in existing literature, this chapter aims to develop a real-time speech emotion

 DOI: 10.5281/zenodo.11049046

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

recognition system capable of analysing emotional cues as they unfold in natural conversation. The proposed system encompasses key components such as voice activity detection, speech segmentation, signal pre-processing, feature extraction, emotion classification, and statistical analysis of emotion frequency. Experimental evaluations demonstrate high accuracy rates, validating the effectiveness of the system in accurately identifying and categorizing emotions from continuous speech data.

Additionally, the chapter explores the practical application of the developed system in online learning environments, showcasing its potential to personalize online courses based on students' emotional responses to course content. Overall, this research contributes to the advancement of emotion recognition technology by proposing a real-time speech-based approach and demonstrating its potential utility in practical settings such as online education [7].

Sonali T. Saste et.al (April 2017). In recent years, there has been a surge of interest in emotion recognition from speech within the domain of human-computer interaction. Researchers have explored various methodologies and systems aimed at accurately identifying emotions expressed in speech signals. A notable advancement in this field is the development of language-independent systems, allowing for emotion recognition across different languages. These systems typically rely on databases of emotional speech samples for feature extraction, with algorithms such as Mel-frequency cepstral coefficients (MFCC) and Discrete Wavelet Transform (DWT) being commonly utilized. Classification of emotions is often achieved through machine learning techniques such as SVM classifiers. One intriguing application of speech emotion recognition is its integration into security systems, such as ATM security, where recognized emotions can be used to enhance security protocols. This ongoing research underscores the importance of emotion recognition in improving human-computer interaction and highlights its potential for diverse practical applications [8].

Ryota Sato et.al (Nov 2020). Speech Emotion Recognition (SER) has emerged as a significant challenge within the realm of human-computer interaction. Conventional SER methods have typically been constrained by the limitations of existing speech-emotional databases, which often assign a single emotion label to each utterance. However, human speech is inherently complex, often conveying multiple emotions simultaneously with varying intensities.

Recognizing this nuanced interplay of emotions is crucial for achieving a more natural and accurate SER. To address this gap, researchers have begun exploring the creation of emotional speech databases that incorporate labels for multiple emotions and their intensities within each utterance. By capturing the richness and complexity of human emotional expression, these databases hold the potential to significantly enhance the sophistication and realism of SER models. Moreover, advancements in experimental techniques, such as the extraction of emotional segments from existing video works, have enabled researchers to create more comprehensive and diverse emotional speech databases for training and evaluation purposes. As a result, the field of SER is evolving rapidly, with a growing emphasis on capturing and recognizing the multifaceted nature of emotions in human speech [9].

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

Vinita Chugh et.al (Oct 2021). The significance of recognizing emotions conveyed through speech highlights its natural and essential role in human communication. It discusses the challenges associated with speech emotion recognition, emphasizing the complexity of human emotions and the difficulties in differentiating between them, particularly in practical and uncontrolled conditions. The abstract also addresses limitations in existing emotional speech datasets, pointing out the scarcity of data with emotional classifications.

These challenges underscore the laborious nature of speech emotion recognition, necessitating comprehensive approaches that consider various factors such as environmental noise, context, and individual intentions. Despite these challenges, the abstract suggests that advancements in emotion detection techniques and the exploration of predictive feature sets could offer promising avenues for enhancing the accuracy and effectiveness of speech emotion recognition systems [10].

IV. IMPLEMENTATION

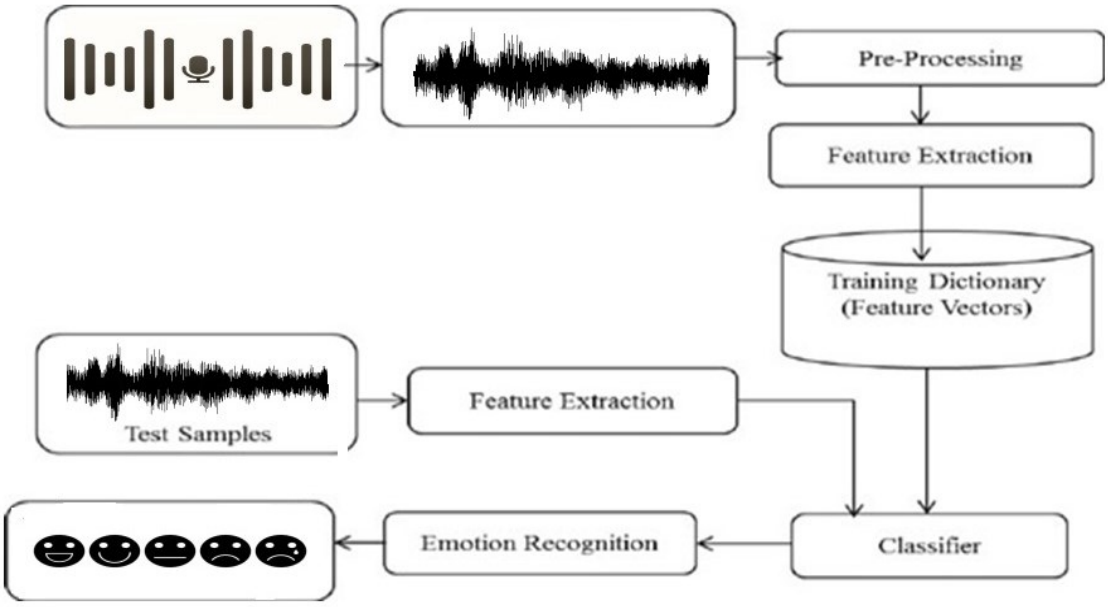


Figure 2: Implementation of the System

1. **Import Libraries:** To import libraries for speech emotion detection, use 'librosa' for audio analysis, 'NumPy' for numerical data handling, and 'scikit-learn' for machine learning tools. For deep learning, consider 'TensorFlow' or 'PyTorch' with 'Keras' API. Visualize using 'Matplotlib'.

2. **Load Dataset:** The RAVDESS dataset contains 1440 audio files featuring performances by 24 professional actors (12 male, 12 female) delivering lexically matched statements in a neutral North American accent. It covers 8 emotions. Each file is uniquely labelled with a 7-part numerical identifier, encoding actor, statement, emotion, intensity, repetition, gender, and trial number.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

This dataset is invaluable for research in emotion recognition and natural language processing, enabling in-depth analysis of emotional expression in speech. Filename example: 03-01-06-01-02-01-12.wav.

3. **Data pre-processing:** Development of full preliminary data. It is a process that increases knowledge by creating new information from existing content. This includes small changes to the data to improve the performance of the model. It is worth noting that composite data generated in MFCC speech can improve speaker recognition through transfer learning. Some methods of data improvement include analyzing the quality of the data by injecting noise into the data set.

4. **Features Selection and Extraction:** The MLP classifier considers three features, which is an important step after selecting suitable data to train the algorithm. Although there is a tendency to include many things, especially inclusive ones, doing so can affect the quality of a traditional product. Traditional algorithms are valued for their speed and efficiency, even if they sacrifice some accuracy compared to more complex methods. In addition to the two functions popular in speech recognition – Mel Frequency Cepstral Coefficients (MFCC) and Mel Frequency Cepstral Coefficients (MFC) – there is a third function called Chroma, which provides music. A compressed representation of the tonal content of the signal is exploited. Chroma is a new addition to the space and is introduced here to evaluate its impact on the model's performance.

5. **Classifier:** After preprocessing the features to ensure uniformity, a machine learning model, like a Multilayer Perceptron (MLP) classifier, is chosen and trained on the labelled data. The model's hyperparameters are tuned using a validation set to optimize performance. Evaluation metrics such as accuracy are then used to assess the model's effectiveness. Once satisfied with the model's performance, it can be deployed for use in various applications, providing insights into the emotional content of spoken language.

6. **Result:**

```python
#Emotions in the RAVDESS dataset to be classified Audio Files based on .
emotions={
  '01':'neutral',
  '02':'calm',
  '03':'happy',
  '04':'sad',
  '05':'angry',
  '06':'fearful',
  '07':'disgust',
  '08':'surprised'
}
#These are the emotions User wants to observe more :
observed_emotions=['calm', 'happy', 'fearful', 'disgust']
```

Python

Result 1: Total Eight labeled emotions

ISSN (Online): 2583-648X

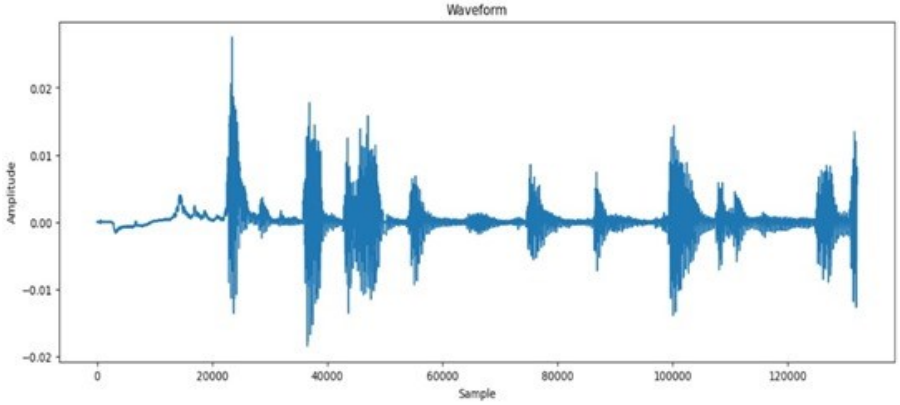
IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in



Result 2: Wavefrom of The Speech Input

```
## Appying extract_feature function on random file and then loading model to predict the result
import librosa
import numpy as np

# Assuming extract_feature is defined and imported correctly

file = 'output10.wav'
# Load the audio file and extract features
data, sr = librosa.load(file, sr=16000)  # Load audio with a specific sampling rate
new_feature = extract_feature(file, mfcc=True, chroma=True, mel=True)

# Only consider the first returned value
ans = np.array(new_feature).reshape(1, -1)

# Predict the emotion label
predicted_label = Emotion_Voice_Detection_Model.predict(ans)
print(predicted_label)
```

[90] Python

```
['calm']
```

Result 3: Final Output As Predicted Emotion

V. ADVANTAGES AND DISADVANTAGES

Advantages:

1. Improving human-machine interaction: Helping virtual assistants and entertainment platforms improve user experience and encourage natural interactions.
2. Depression Research: Monitoring mental status to assist doctors and researchers with early intervention and treatment.
3. Cross-cutting applications: A versatile tool for drawing inferences from verbal data for education, customer needs analysis, and healthcare.
4. Real-time analysis: Improve user experience and security by supporting the real-time performance of applications.

Disadvantages:

1. Data dependency: Truth depends on the presence and quality of information about emotions, thus affecting the functioning of the body.
2. Variability in Emotional Expression: Establishing good international standards is difficult due to cultural and personal teachings and more research is needed to update them.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 2 | April 2024

www.ijirid.in

VI. CONCLUSION

The Speech Recognition (SER) using machine learning project represents a major effort at the intersection of artificial intelligence and human perception. Emotions are the basis of human interaction and are intertwined with speech, so knowing emotions is important for effective communication. We aim to bridge the gap between human thought and technology by using machine learning algorithms and computational techniques. The paper addresses the growing need for the use of emotions, where machines can not only understand human emotions but also respond to them instantly. The motivation behind emotion recognition (SER) machine learning has many real-life applications, including human-computer interaction, user sentiment analysis products, and psychological testing. By creating powerful SER systems, we pave the way for more intuitive and responsive technologies that engage with their users on a deeper level. Looking to the future, the integration of cognitive technology is expected to bring about changes in many areas, affecting the amount of time people use computers, where technology truly understands and enables human emotions.

REFERENCES

[1] Utkarsh Garg, Sachin Agarwal, Shubham Gupta, Ravi Dutt and Dinesh Singh, “Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma,” 12th International Conference on Computational Intelligence and Communication Networks, Nov 2020.

[2] Anguish Gan grade, Shalini Singhal, “A Research of Speech Emotion Recognition Based on CNN Network,” SKIT Research Journal, VOLUME 12, ISSUE 1, July 2022.

[3] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, “Speech based Emotion Recognition using Machine Learning,” 3rd International Conference on Computing Methodologies and Communication (ICCMC), March 2019.

[4] S. G. Shaila, A. Sindhu, L. Manish, D. Shivamma, and B. Vaishali, “Speech Emotion Recognition Using Machine Learning Approach,” ICAMIDA 2022, ACSR 105, pp. 592–599, May 2023.

[5] T. Kishore Kumar, “Stressed Speech Emotion Recognition using feature fusion of Teager Energy Operator and MFCC,” IEEE – 40222 8th ICCCNT, July 2017.

[6] Chen Caihua, “Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM,” 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), July 2019.

[7] Husbaan I. Attar, Nilesh K. Kadole, Omkar G. Karanjekar, Devang R. Nagarkar, Prof. Sujeet and More, “Speech Emotion Recognition System Using Machine Learning,” International Journal of Research Publication and Reviews, Vol 3, no 5, pp 2869-2880, May 2022.

[8] Sonali T. Saste, Prof. S. M. Jagdale, “Emotion Recognition from Speech Using MFCC and DWT for Security System,” International Conference on Electronics, Communication and Aerospace Technology ICECA, April 2017.

[9] Ryota Sato, Ryohei Sasaki, Norisato Suga, Toshihiro Furukawa, “Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition,” Proceedings of the 23rd Conference of the Oriental COCOSDA, Yangon, Myanmar, Nov 2020.

[10] Vinita Chugh, Shivanghee Kaw, Surabhi Soni, Varsha Sablani & Rupali Hande, “Speech Emotion Recognition System Using MLP,” 2021 JETIR October 2021, Volume 8, Issue 10 www.jetir.org (ISSN-2349-5162), Oct 2021.

